© М. Ю. Мухин
УрФУ, г. Екатеринбург

СТАТИСТИЧЕСКИЙ ПОТЕНЦИАЛ ЛЕКСИКОГРАФИЧЕСКИХ БАЗ ДАННЫХ УРАЛЬСКОЙ СЕМАНТИЧЕСКОЙ ШКОЛЫ[1]

Лексикографическим базам данных проблемной группы «Русский глагол» и Уральской семантической школы недавно исполнилось 20 лет. Первая база на основе СУБД FoxPro была создана на кафедре современного русского языка в 1993 г. для работы с толковым идеографическим словарем русских глаголов. В среде, работающей под DOS, были созданы процедуры, обеспечивающие удобный на тот момент ввод данных. Согласно требованиям идеографической организации, данные были размещены в связанных друг с другом таблицах, отражающих уровни семантической иерархии. После заполнения контентом словаря глаголов в базе середины 1990-х гг. было около 6,5 тыс. записей. Отдельную проблему в то время составлял экспорт информации в текстовые процессоры и издательские системы в силу разницы в форматах, а также технические неприятности с рядом внешне совпадающих латинских и кириллических символов.

База данных словаря глаголов в своем развитии прошла ряд этапов в связи с совершенствованием инструментальных возможностей, а также изменением самого контента: появлением дополнительных типов информации и выделения отдельных записей для членов видовых пар — глаголов совершенного вида, которые привязаны в словаре к вокабулам (глаголам несовершенного вида). Впоследствии по модели глагольного словаря была спланирована архитектура данных для идеографических словарей существительных и прилагательных, а также для словаря-тезауруса русских синонимов. Эти базы сделаны уже на основе СУБД MS Access в силу ее доступности для непрофессиональных пользователей и интегри-

[1] Исследование выполнено при финансовой поддержке РГНФ (проект 13-04-00322 «Русская лексика как межчастеречная система: полное идеографическое описание в лексикографических параметрах») и при поддержке средств, полученных из Программы повышения конкурентоспособности (номер соглашения 02.A03.21.0006).

рованности в пакет MS Office, обеспечивающей элементарный обмен данных между основными интересующими нас приложениями: Access — Word — Excel.

В 2012–2013 гг. в рамках работы над проектом Универсального идеографического словаря, который в будущем должен содержать описание слов всех частей речи, включая служебные, были проведены слияние данных разных словарей в одну базу и разработка удобной формы для работы с тезаурусом. Таким образом, новая база данных объединила идеографические структуры и словники четырех словарей: 1) Большой толковый словарь русских глаголов: Идеографическое описание. Синонимы. Антонимы. Английские эквиваленты; 2) Большой толковый словарь русских существительных: Идеографическое описание. Синонимы. Антонимы; 3) Словарь-тезаурус русских прилагательных, распределенных по тематическим группам; 4) Словарь-тезаурус синонимов русской речи. Для более точного и полного формирования словника из соответствующих словарных статей были выделены объекты описания, не являющиеся заголовочными словами (вокабулами). Это словообразовательные и фонетические варианты, видовые пары глаголов, производные существительные, обозначающие лиц женского пола, и т. п. Каждая подобная лексема получила отдельную запись и соотнесена в базе данных с основной вокабулой, организующей словарную статью.

Кроме того, сопоставление с электронными версиями существующих толковых и частотных словарей выявило также частотные лексемы, для которых в Новом частотном словаре русской лексики под ред. О. Н. Ляшевской и С. А. Шарова (URL: http://dict.ruslang.ru/freq.php) установлен показатель *ipm* не менее 3 и которые не были ранее семантизированы в идеографических словарях группы «Русский глагол». Эти лексемы включены в новый словник. В словник вошли также частотные неоднословные лексические единицы с семантикой наречий, вводных слов, предлогов, союзов и частиц. В итоге стартовый вариант базы данных, рассчитанный на создание Универсального словаря и разработку др. проектов, включает 96 966 записей — ЛСВ слов разных частей речи. Из них от словаря глаголов унаследовано 10 443 записи, от словаря существительных — 14 898, от словаря прилагательных — 23 048, от тезауруса синонимов — 42 693. К этому в базу добавлено

5 884 записей — новых ЛСВ. Специфика разных словарей стала причиной количественной асимметричности слов разных частей речи (особенно в плане ЛСВ имен прилагательных) и совпадения ряда значений в тезаурусе синонимов и др. словарях. Эти проблемы должны быть преодолены в процессе окончательной выборки материала для Универсального словаря.

Под руководством проф. Л. Г. Бабенко значительно изменена и дополнена синоптическая схема сводного тезауруса. Выявлены различия между структурами разных словарей, соотнесены рубрикации, произведена переиндексация основной части тезаурусов. Приняты решения по соединению и, наоборот, разделению ряда словарных групп, а также по устранению логико-понятийных нестыковок. На сегодняшний день эта структура объединяет в базе 962 записи — наименования денотативных сфер, подсфер и реальных лексических групп. Схема имеет 6 уровней и проявляет наибольшую дробность в сфере «3. Живая природа» (например, группа «3.2.2.3.2.3. Грызуны»).

Кроме новых лексикографических задач, которые позволяет решить сегодняшняя структура данных, мы видим у нее базовые статистические возможности, обусловленные самим контентом: семантической классификацией и полями, которые в основном соотносятся со словарными зонами словарей. Приведем из этих возможностей только наиболее очевидные:

— количественное соотношение денотативных сфер и групп, в том числе в синонимическом словаре: выявление специфики языковой картины мира на лексикографическом материале;

— стилистическое распределение внутри тезауруса (пометы): выявление функционально-стилистических, эмоциональных, оценочных приоритетов в разных денотативных классах, а также у слов, вступающих в синонимические отношения;

— распределение прямых и переносных значений: выявление образно-метафорических приоритетов;

— соотношение описываемых слов и идентификаторов, употребляемых в толкованиях: автоматизированное выявление гипо-гиперонимических отношений в лексике;

— частеречное соотношение и распределение грамматических категорий (например, глагольного вида, рода и числа существи-

тельных и т. д.): выявление грамматических особенностей денотативных классов.

Как видим, разработка даже чисто статистических аспектов баз данных Уральской семантической школы имеет сегодня большой научный потенциал, что, конечно, отразится в новых исследованиях на лексикографическом материале.

© М. Э. Рут
УрФУ, г. Екатеринбург

К ВОПРОСУ ОБ ИДЕОГРАФИЧЕСКОМ СЛОВАРЕ АНТРОПОНИМОВ

Лексикографическая практика в области антропонимики достаточно монотонна: обычно это алфавитный перечень личных имен, сопровождаемый теми или иными лингвистическими комментариями (особенности склонения, наличие гипокористик, этимологическая справка и т. п.). Именно таким образом построены наиболее известные словари: словарь Н. А. Петровского [1], А. В. Суперанской [2] и др. Содержащаяся в таких лексиконах информация полезна и интересна широкому кругу читателей, и они пользуются заслуженным спросом.

В этой же лексикографической традиции оформляются и словари прозвищ, самым капитальным из которых следует признать словарь В. М. Мокиенко, Х. Вальтера [3], за тем исключением, что здесь наиболее важной информацией служит справка о мотивационной модели именования (если в таковой есть необходимость). Между тем именно прозвища, а также фамилии наводят на мысль о возможности идеографической интерпретации антропонимического материала в рамках лексикографической практики.

Если опираться на основной принцип идеографического словаря — построение материала не в порядке наименований, а исходя из логики заложенных в них смыслов, то для антропонимов есть два пути лексикографического развития. Первый — это создание словаря, где антропонимы скомпонованы по группам антропонимных референтов, т. е. в основу словарных статей положены социальные термины, характеризующие носителей антропонимов. Такой принцип был бы весьма интересен для словаря фамилий: учет наиболее